粗集模型的特征函数表示*

赵晓雨[1]，雷晓蔚[2]
(1. 重庆文理学院 应用技术学院；2. 重庆文理学院 物理与信息工程系，重庆 402160)

摘 要 粗糙集理论是处理不确定知识的一种工具，已在人工智能与知识发现、模式识别与分类、数据挖掘与故障检测等方面得到了较好应用。由于粗糙集在理论和应用两个方面的迅速发展，粗集模型得到拓广。本文研究粗集模型的特征函数表示形式，这种表示形式具有一般性，可以统一各种推广模型。粗集理论的核心是一对非数值型算子，即上下近似算子。粗集理论中的上下近似算子与证据理论中的一对数值算子——似然函数和信任函数有密切关系，为此作者研究了粗糙集与证据理论的关系。

关键词 粗集；上下近似算子；信任函数；特征函数；证据理论
中图分类号：TP18 文献标识码：A 文章编号：1672-6693(2007)04-0054-04

粗糙集理论作为一种处理不精确(imprecise)、不一致(inconsistent)、不完整(incomplete)等各种不完备信息有效的工具，作为一种数据分析处理理论，于 1982 年由波兰科学家 Z. Pawlak 创立[1]。进入网络信息时代，随着计算机技术和网络技术的飞速发展，使得各行业领域的信息急剧增加，数据挖掘(Data Mining)和知识发现(KDD)技术应运而生。粗糙集理论由于在数据挖掘方面的应用而备受关注[2-3]。最近几年，粗糙集理论的应用研究得到了长足发展[4]，出现了许多粗糙集的扩展模型。粗集模型的推广一直是粗集理论研究的主流之一[5-8]。例如，Yao[9] 把粗集的上下近似推广到最一般的二元关系上，并讨论了在不同关系下粗糙集模型的性质；张文修教授[10-11]等给出了一类基于随机集的粗集模型，又在文献[12]中把这种模型解释为间接学习的模型。本文用集合的特征函数的方法把粗集定义的多数形式统一起来，给出了粗集较为一般的形式，改进了以前文献的结果，为粗集的研究提供一种有效方法。最后还讨论了粗集上下近似与证据理论的关系。

本文在讨论粗集时所涉及的论域既可以是有限集也可以是无限集，为了使得粗集模型有较大的概括性，笔者既讨论了一个论域上的粗集，又在两个论域 U，W 上考虑问题，设 R 为 U 到 W 上的二元关系。

1 一个论域上的粗集上下近似的定义

设 U 是论域(有限无限均可)，对于一个论域的情况，Yao[9] 给出了最为一般的粗糙集的定义，其上下近似的定义是通过邻域来实现的，设 $r(x)=\{y\mid y\in U, xRy\}$，$r(x)$ 称为元素 x 的 R 邻域(或简称为 x 的邻域)。设 X 为 U 的子集，X 的上下近似(分别记为 $\overline{R}X$，$\underline{R}X$)的定义为

$$\overline{R}X=\{x\mid x\in U, r(x)\cap X\neq\varnothing\},\ \underline{R}X=\{x\mid x\in U, r(x)\subseteq X\}。$$

设 $P(U)$ 表示 U 的幂集，与文献[13]一样，若 X 是 U 的子集，X 既表示子集，又表示 X 的特征函数，其区别可以通过上下文来区分，以下定理是粗集上下近似的特征函数的表述形式。

定理 1 设 U 为论域，R 为 U 上的二元关系，$X\in P(U)$，子集 X 的上下近似用特征函数可表述如下：

1) X 的上近似 $\overline{R}X\in P(U)$；$(\overline{R}X)(x)=\bigvee\limits_{y\in U}(R(x,y)\wedge X(y))$，$\forall x\in U$；

2) X 的下近似 $\underline{R}X\in P(U)$；$(\underline{R}X)(x)=\bigwedge\limits_{y\in U}((1-R(x,y))\vee X(y))$，$\forall x\in U$。

这里 $\wedge$，$\vee$ 分别表示取小、取大运算。

---

* 收稿日期：2007-03-01
资助项目：重庆市教育委员会科学技术研究项目(No. KJ061208)
作者简介：赵晓雨(1966-)，男，山东苍山人，副教授，硕士，研究方向为人工智能、数据挖掘。

证明 仿照文献[13]定理1的证明进行。 证毕。

2 随机集的粗集模型

张文修教授[10-11]等为了研究随机集的粗集模型,给出了集值映射的上下近似的定义,后来又在文献[12]中把该模型解释成间接学习的理论模型,并探索了间接学习的质量问题,其核心都是集值映射上下近似理论,为方便笔者给出集值映射上下近似的定义,当论域为有限集合时定义见文献[12],这里不假定论域是有限集合。

定义1 设 U,W 为两个论域,$P(W)$ 为 W 的幂集,$F: U\to P(W)$ 是集值映射,对任意的 $X\subseteq W$,记

$$\underline{F}X=\{x\mid x\in U,F(x)\subseteq X\}\ \text{及}\ \overline{F}X=\{x\mid x\in U,F(x)\cap X\neq\varnothing\}$$

则 $\underline{F},\overline{F}$ 分别称为 X 关于集值映射 F 的下近似和上近似,而分别称 $\underline{F},\overline{F}: P(W)\to P(U)$ 为下近似算子与上近似算子。

正如文献[13]指出,集值映射 $F: U\to P(W)$ 与 U 到 W 的二元关系 R 可以相互唯一确定。对应关系可以通过"xRy 当且仅当 $y\in F(x)$"来实现。笔者指出用特征函数,集值映射 F 的上下近似可以表述如下。

定理2 设 U,W 为两个论域,$P(W)$ 为 W 的幂集,$F:U\to P(W)$ 是集值映射,其相应的 U 到 W 的二元关系为 R,则对任意的 $X\subseteq W$,有

1) X 的上近似 $\overline{F}X\in P(U)$;$(\overline{F}X)(x)=\bigvee\limits_{y\in W}(R(x,y)\wedge X(y)),x\in U$;

2) X 的下近似 $\underline{F}X\in P(U)$;$(\underline{F}X)(x)=\bigwedge\limits_{y\in W}((1-R(x,y))\vee X(y)),x\in U$。

这里 $\wedge,\vee$ 分别表示取小和取大运算。

证明 由文献[13],U 到 W 上的二元关系 R 定义为 xRy 当且仅当 $y\in F(x)$。

1) 若 $\overline{F}X(x)=1$,即 $x\in\overline{R}X$,由定义1,$F(x)\cap X\neq\varnothing$,于是存在元素 $a\in F(x)\cap X$,即 xRa 且 $a\in X$,因此 $R(x,a)=X(a)=1$,这样 $\bigvee\limits_{y\in W}(R(x,y)\wedge X(y))=1$,即证明了:若 $\overline{R}X(x)=1$,则 $\bigvee\limits_{y\in W}(R(x,y)\wedge X(y))=1$。

反之,若 $\bigvee\limits_{y\in U}(R(x,y)\wedge X(y))=1$,则存在 $x\in U,y\in W$,使得 $R(x,y)\wedge X(y)=1$,即 $R(x,y)=X(y)=1$,于是 $y\in F(x)\cap X$。因此,$y\in F(x)\cap X$,这样 $F(x)\cap X\neq\varnothing$,$x\in\overline{F}X$,即 $x\in\overline{F}X$ 当且仅当 $\bigvee\limits_{y\in W}(R(x,y)\wedge X(y))=1$ 得证,因此定理2的结论1)成立。

2) 若 $\underline{R}X(x)=1$,即 $x\in\underline{R}X$。注意到,如果 $R(x,y)=1$,这时 $y\in F(x)$,由下近似的定义,$y\in X$,即 $X(y)=1$;如果 $R(x,y)=0$,有 $1-R(x,y)=1$。总有 $\bigwedge\limits_{y\in W}((1-R(x,y))\vee X(y))=1$,即若 $\underline{R}X(x)=1$,可以推出 $\bigwedge\limits_{y\in W}((1-R(x,y))\vee X(y))=1$ 得证。

反之,设 $x\in U$,若 $\bigwedge\limits_{y\in W}((1-R(x,y))\vee X(y))=1$,则对于任意 $y\in W$ 有 $(1-R(x,y))\vee X(y))=1$,即 $1-R(x,y)=1$ 或 $X(y)=1$,这样,$R(x,y)=0$ 或 $X(y)=1$。若 $y\in F(x)$,则 $R(x,y)=1$,且 $X(y)=1$,即 $x\in\underline{R}X$。因此有 $\underline{R}X(x)=1$ 当且仅当 $\bigwedge\limits_{y\in W}((1-R(x,y))\vee X(y))=1$,即定理2的结论2)成立。

证毕

当 U,W 均为两个有限论域时,这时,设 U 到 W 的二元关系 R 可以由其相应的关系(布尔)矩阵来表述,这样定理2的结论1)就是文献[13]的定理1。因此,本文改进了文献[13]的结果。

3 两个论域上粗集的定义

受定理1及定理2的启发,可以用特征函数给出两个论域上粗集的一般定义。

定义2 设 U,W 为两个论域,R 为 U 到 W 的任意二元关系,对任意的 $X\subseteq W$,定义

1) X 的上近似 $\overline{R}X\in P(U)$,$(\overline{R}X)(x)=\bigvee\limits_{y\in W}(R(x,y)\wedge X(y)),x\in U$;

2) X 的下近似 $\underline{R}X\in P(U)$,$(\underline{R}X)(x)=\bigwedge\limits_{y\in W}((1-R(x,y))\vee X(y)),x\in U$。

这里 $\wedge,\vee$ 分别表示取小、取大运算。这时分别称 $\overline{R},\underline{R}: P(W)\to P(U)$ 为粗集的上下近似算子,称 $(\overline{R}X,\underline{R}X)$

为 X 的粗集。

由定理 1,容易看出,当 $U = W$ 时,定义 2 就是 Yao[9] 给出的一个论域时的粗集定义。只需注意到集值映射 $F: U \to P(W)$ 与 U 到 W 的二元关系 R 可以相互唯一确定[13],由定理 2,定义 2 就是张文修教授[10-12] 定义的两个论域的粗集,只不过这里不要求论域是有限集合。因此定义 2 给出了粗集较为一般的形式。

4 粗集上下近似与证据理论的关系

本节笔者来研究粗集上下近似与证据理论的关系。Dempster 和 Shafer 在上世纪 60 年代建立的证据理论可以用来处理由不知道引起的不确定性,它是概率论的进一步扩充,该理论主要是通过概率分配函数、信任函数及似然函数来表述的,由莫比乌斯(Mobius) 变换知道概率分配函数 m、信任函数 Bel 及似然函数 Pl 三者可以相互唯一确定。而粗集理论的主要思想是不精确的概念如何用可利用的知识库中的知识来近似地描述,其核心是一对非数值型算子,即上下近似算子。粗集理论中的上下近似算子与证据理论中的一对数值算子 —— 似然函数和信任函数有密切关系。

容易看出,证据理论的分配函数、信任函数及似然函数与粗集的等价类、下近似及上近似有极为类似的关系,这可由以下对应关系看出(设 U 为有限论域,且 R 为 U 上的等价关系,R 确定的商集为 U/R)m,Bel, $Pl: P(U) \to [0,1]$,m 按照

$$m(X) = \begin{cases} |X|/|U|, & \text{当 } X \in U/R \text{ 时} \\ 0, & \text{当 } X \notin U/R \text{ 时} \end{cases}$$

证据理论:	m	$Bel: Bel(X) = \sum_{Y \subseteq X} m(Y)$	$Pl: Pl(X) = \sum_{Y \cap X \neq \varnothing} m(Y)$
	↕	↕	↕
粗集理论:	U 的划分	下近似 $\underline{R}X = \bigcup_{[x] \subseteq X} [x]$	上近似 $\overline{R}X = \bigcup_{[x] \cap X \neq \varnothing} [x]$

对于集值映射有类似的关系,设 $F: U \to P(W)$ 是集值映射,只要满足 $F(x) \neq \varnothing$, $\forall x \in U$,通过映射 $j: P(W) \to P(U)$

$$j(Y) = \begin{cases} [x], & \text{如果存在 } x \text{ 使 } F(x) = Y \\ \varnothing, & \text{如果对于任意的 } x \text{ 均有 } F(x) \neq Y \end{cases}$$

及 $m(Y) = \begin{cases} |j(Y)|/|U|, & \text{当 } j(Y) \in U/\sim \text{ 时} \\ 0, & \text{当 } j(Y) \notin U/\sim \text{ 时} \end{cases}$,$Y \in P(W)$ 有

证据理论:	m	$Bel: Bel(X) = \sum_{Y \subseteq X} m(Y)$	$Pl: Pl(X) = \sum_{Y \cap X \neq \varnothing} m(Y)$
	↕	↕	↕
集值映射粗集模型:	U 的划分	下近似 $\underline{F}X = \bigcup_{F(x) \subseteq X} [x]$	上近似 $\overline{F}X = \bigcup_{F(x) \cap X \neq \varnothing} [x]$

5 结束语

粗糙集理论认为知识的粒度性是造成使用已有知识不能精确地表示某些概念的原因,通过引入不可区分关系作为粗糙集理论的基础,并在此基础上定义了上下近似等概念,它是粗集理论的关键概念,上下近似与论域及其定义在其上的二元关系密切相关。在研究方法上,笔者使用特征函数的方法,与集合论的方法形成互补,为研究问题带来很大的方便。

数据挖掘是当今人工智能、数据库和统计学等领域的前沿课题,据预测在未来数年中数据挖掘会有更广泛的应用。与之相适应,粗糙集理论近年来获得了飞速发展。经典粗糙集模型采用等价关系(大多数情况是相等关系)作为基础,但是由于这不能完全满足应用的需要,因此提出了不少新的模型和扩展。本文使用的特征函数的方法使得粗集模型有了一个统一的形式,采用这种表示方法可以使许多问题得到简化,这就为粗集的研究提供了一种行之有效的方法。

参考文献：

[1] PAWLAK Z. Rough sets [J]. International Journal of Information and Computer Science,1982,11(5):341-356.

[2] WAIYAMAI K,LAKHAL L. Knowledge Discovery from very Large Databases Using Frequent Concept Lattice[C]. Heidelbeng: Springer Berlin,2000,1810:437-445.

[3] 丁保淼,张运陶. 应用粗糙集理论分析识别岩石种类的因素[J]. 西华师范大学学报(自然科学版),2005,26(2):161-165.

[4] 胡可云,陆玉昌,石纯一. 粗糙集理论及其应用进展[J]. 清华大学学报,2001,41(1):64-68.

[5] YAO Y Y,LINGRAS P J. Interpretations of Belief Functions in the Theory of Rough Sets[J]. Information Sciences, 1994,104(1-2):81-106.

[6] 刘清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001.

[7] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001.

[8] 雷晓蔚. 粗集理论的矩阵方法[J]. 计算机工程与应用,2006,42(17):73-75.

[9] YAO Y Y. Constructive and Algebraic Methods of the Rough Sets [J]. Information Sciences,1998,109(1-4):21-47.

[10] 张文修,吴伟志. 基于随机集的粗集模型(I)[J]. 西安交通大学学报,2000,34(12):75-79.

[11] 张文修,吴伟志. 基于随机集的粗集模型(Ⅱ)[J]. 西安交通大学学报,2001,35(4):425-429.

[12] 米据生,张文修. 基于粗集的间接学习[J]. 计算机科学,2002,29(6):96-97.

[13] 刘贵龙. 基于两个集合上粗集模型的算法实现[J]. 计算机科学,2006,33(3):181-184.

Characteristic Function Representation of Rough Sets

ZHAO Xiao-yu[1], *LEI Xiao-wei*[2]

(1. College of Applied Science and Technology;

2. Dept. of Physics and Information Engineering, Chongqing University of Arts and Sciences, Chongqing 402160, China)

Abstract: Rough set theory is a new tool dealing with uncertainty. We have found its applications in many areas such as artificial intelligence (AI), knowledge discovery (KDD), pattern recognition and classification, and data mining and fault diagnostication. Various generalization of rough set in lower and upper approximation is given due to the development of the rough set theory and its application. The paper studies the characteristic function representation of rough set. This representation is universal. Unified characteristic function form of lower and upper approximation is given. The core of rough set theory is a pair of non-numerical operators, i. e. lower and upper approximation operators. Lower and upper approximation operators have close relation with likelihood functions and belief functions that are a pair of numerical operators in the Shafer's evidence theory. So it is necessary to investigate the relationship between rough set and Shafer's evidence theory.

Key words: rough set; lower and upper approximation operators; belief functions; characteristic function; evidence theory.

(责任编辑　游中胜)